Final Project Proposal

Students:

Dan Shamia (208004119) and Daniel Halperin (207826314)

Topic:

Data-Centric Analysis of Hallucination and Confidence Calibration in LLMs

Proposal:

The reliability of Large Language Models (LLMs) remains a critical bottleneck for their deployment in high-stakes environments. While earlier work focused on categorizing hallucinations, as seen in *"Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models",* recent research has shifted toward understanding the mechanisms driving them. Notably, the study *"Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?"* demonstrated that fine-tuning on new knowledge can paradoxically encourage hallucinations, suggesting a complex relationship between data exposure and factual grounding. Furthermore, *"Hallucination is Inevitable: An Innate Limitation of Large Language Models"* argues that hallucination is an inherent limitation of statistical modeling, necessitating better confidence calibration as proposed in recent comprehensive reviews like *"A Survey on Hallucination in Large Language Models"*.

In this project, we investigate how the frequency and distribution of factual training data directly impact both the accuracy and the internal confidence of modern language models. We will utilize the FEVER dataset (Fact Extraction and VERification) to construct controlled experimental splits representing high-exposure (common knowledge) and low-exposure (long-tail) facts.

We will perform Supervised Fine-Tuning (SFT) on state-of-the-art open-weights models, specifically Mistral-7B and the Llama-3 family (8B). By manipulating the training data distribution, we aim to isolate the effect of "data frequency" on the model's propensity to hallucinate. Evaluation will go beyond standard accuracy and F1 scores to include confidence calibration metrics (such as Expected Calibration Error), specifically analyzing "high-confidence errors"-instances where the model is confidently wrong.

The primary innovation of this project is a rigorous, controlled data-centric analysis. Rather than proposing a new architecture, we aim to provide empirical evidence linking specific training data distributions to the emergence of hallucinations and miscalibration in modern LLMs.